Experimenting Texture Similarity Metric STSIM for Intra Prediction Mode Selection and Block Partitiong in HEVC

Karam Naser, Vincent Ricordel, Patrick Le Callet

Experimenting Texture Similarity Metric STSIM for Intra Prediction Mode Selection and Block Partitiong in HEVC

Karam Naser, Vincent Ricordel, Patrick Le Callet
LUNAM University, University of Nantes, IRCCyN UMR CNRS 6597
Polytech Nantes, Rue Christian Pauc BP 50609 44306 Nantes Cedex 3, France
karam.naser@univ-nantes.fr
vincent.ricordel@univ-nantes.fr
patrick.lecallet@univ-nantes.fr

***Abstract*—Textures can often be found in large areas of images and videos. They have different spectral and statistical properties as compared to normal (structural) components. Encoding them with ordinary video coders requires higher bit rate and usually results are unsatisfying in perceived quality. Recently, different perceptual tools have been developed to estimate the perceived quality of textures taken into account models of Human visual system. In this paper, we investigate and discuss the practical usability of one of these tools, namely STSIM, as a distortion function for selecting the intra-prediction mode and block partitioning of texture images in HEVC. We experiment few practical implementations to examine its performance compared to default metrics used by HEVC. Experimental results showed that the perceived quality of the decoded textures has been significantly improved specially for stochastic types of textures**

***Index Terms*—Texture Coding; Perceptual Optimization, STSIM**

I. Introduction

Video coding standards, such as HEVC [1], aim at representing video signals with a minimum bitrate at a certain distortion level. The typical distortion measure is based on comparing individual pixels values of the original and distorted signals. This kind of measure, however, does not correctly fit our visual perception, in other words, the amount of distortion that it measures does not proportionally reflect the amount of distortion that we perceive. To overcome this problem, many researchers have introduced various *Perceptual Distortion* metrics in the context of *Perceptual Video Coding*. These metrics are based on models of Human Visual System (HVS). The details of such metrics and coding techniques can be found in [2].

In visual perception, the visual signal can be classified into two components, *Texture* and *Structure*. Textures represent homogeneous areas of video scenes with coherent statistics, whereas Structures represent rather the semantics of the scene. Textures appear in large areas of video streams. They can appear in different forms such as sand, grass, tree leaves, see waves and others. Estimating the amount perceptual distortion in the compressed textures is highly different from structures. The is because the Human Visual System focuses on the semantic meaning of the texture rather than the exact details of each pixel. This fact has been exploited in some approaches of texture coding where the textures are removed at the encoder side and re-synthesized at the decoder side. The texture can be synthesized with few parameters to give approximately the same perceptual quality. Examples of these approaches can be found in [3] and [4].

Perceptual video coding can be implemented in several levels of the coding process. It can be done at a pre-processing level such as removal of high frequency components. It can also be done at the post processing level by optimization the rendering filter. At the encoding level, many other approaches have been developed. Examples of this are Region of Interest (ROI) coding, Rate Quality Optimization, Perceptual Quantization and others. All of these approaches try to optimize the perceptual quality of the decoded videos at a given constraints.

In this paper, we investigate the possible perceptual optimization of texture coding in HEVC framework. We used STSIM [5], as being a perceptual similarity metric designed for textures, for selecting the prediction mode and block partitioning in intra-prediction scheme within a fixed quantization parameter scenario. We implemented this metric and compared its performance with the default metrics in HEVC and also with the well known similarity metric (SSIM).

The rest of the paper is organized as follows: Sec. II gives an overview of HEVC intra-prediction mode and Texture similarity metrics used in this work. Sec. III presents the evaluation procedure we used to analyze each metric. In Sec. IV, the simulation results and discussion is presented with conclusion and possible further research is given in Sec. V.

II. Theoretical Background

A. Intra-Prediction Scheme in HEVC

HEVC encoder starts by dividing video frames into Coding Tree Units (CTUs). Each CTU contains ($M \times M$) Coding Tree Block (CTB) for luminance component and 2 ($M/2 \times M/2$) for chrominance components. The CTB is the basic container of one or multiple Coding Blocks (CB). The encoder assigns first one CB for the CTB and tries encoding different prediction schemes and modes. It chooses the one that minimizes the rate-distortion function. The CB can be partitioned in Quadtree manner to find a better rate-distortion value in a smaller block size.

In intra-prediction scheme, each CB contains either 1 or 4 equal sizes Prediction Blocks (PB). For each PB, the encoder searches for the best prediction mode and partitioning into Transform Blocks (TB). The TB's are converted to a bitstream after applying transform and quantization. In HEVC, there are 35 intra-prediction modes. This include 33 directional prediction, DC prediction and Planer prediction. Due to the high complexity of searching for the best prediction mode and quadtree partitioning, the encoder considers only three most probable modes for its optimization process.

B. Overview of Texture Similarity Metrics

We begin this overview with one simple and effective image quality metric which frequently replaces MSE for different applications. It is known as Structural Similarity Index (SSIM). This metric compares the statistics of two image patches *x* and *y*. It is composed of three comparison terms, namely luminance term ($I_{x,y}$), contrast term ($C_{x,y}$) and structure term ($S_{x,y}$). The luminance term compares the mean values of the two patches, where as the contrast term compares the standard deviations of them. The structure term is the cross correlation coefficient between the patches modified with some correction constant. The similarity index between the two patches is calculated as:

$$q_{SSIM} = (I_{x,y})^{\alpha}(C_{x,y})^{\beta}(S_{x,y})^{\gamma} \tag{1}$$

where α, β and γ are design constants and typically take a value of 1.

SSIM has been used for wide range of applications. In [6], it has been used in template matching for generating the prediction signal. It has been improved using the Complex Wavelet domain (CW-SSIM) [7] instead of the spatial domain. Similarly, the frequency coefficients in this domain have been perceptually weighted to fit the perceptual properties of human vision [8].

These metrics, although being useful similarity metrics, may not be reliable similarity metrics for texture images. One of the possible objection is that they allow implicit pixel by pixel comparison inherited in the cross correlation measure of the contrast term. Due to this, an SSIM based approach for texture similarity has been developed. It is known as Structural Texture Similarity Metric (STSIM)[9]. This metric compares the statistics of images subbands performing the following steps:

- **Subband Decomposition:** It uses the Steerable Pyramid Filter[10] to decompose the image into multiple subbands with different orientations and scales. This decomposition has an interesting property of *Shiftability* which allows subbands to have constant power against input shifts.
- **Measure of Statistical differences in Subbands coefficients:** For each subband, the luminance and contrast terms are calculated. The structure term is replaced by the correlation terms. The correlation terms account for horizontal and vertical autocorrelation of the subband.
  STSIM was also improved in [5] by adding the crossband correlation term for comparing the cross correlations coefficients between adjacent bands.
- **Pooling:** The pooling strategy used for STSIM is similar to SSIM except for the cross correlation term. That is, for each subband m, a multiplicative pooling of the similarity terms is calculated:

  $$q^m(x,y) = (I^m_{x,y})^{\frac{1}{4}}(C^m_{x,y})^{\frac{1}{4}}C^m_{xy}(1,0)^{\frac{1}{4}}C^m_{xy}(0,1)^{\frac{1}{4}}$$

  where $C^m_{xy}(1,0)$ and $C^m_{xy}(0,1)$ are the horizontal and vertical cross correlation terms respectively.
  After this, the overall similarity index is calculated by accumulating this term and adding the crossband correlation terms. Let N_b equals to the number of subbands and N_c is the number of crossband correlations, the similarity index is defined as follows:

  $$q_{STSIM} = \frac{\sum_{N_b} q^m(x,y) + \sum_{N_c} C^{m_i,n_i}_{x,y}(0,0)}{N_b + N_c} \tag{2}$$

  where this term $C^{m_i,n_i}_{x,y}(0,0)$ is the cross correlation term which compares the cross correlation between the subbands m and n of x and y.

III. Performance Evaluation

We evaluated respectively the use of SSIM and STSIM metrics as distortion measure inside HEVC for texture coding. Since q_{SSIM} and q_{SSTIM} assess the similarity between two images with a range between $[0, 1]$, we consider as distortion:

$$D_{SSIM} = 1 - q_{SSIM} \tag{3}$$

$$D_{STSIM} = 1 - q_{STSIM} \tag{4}$$

We replaced the distortion measure in HEVC with these two distortions respectively and evaluated the performances of each of them. The details of the experiments and evaluations are given in the following subsections.

A. Framework and Setup

We used the HEVC Modeling software HM9.0 [11] as a host encoder. We chose the Brodatz texture images from USC Viterbi texture dataset [12] in our experiments. These are 13 textures all in gray level images with dimensions 512x512. We converted these textures into YUV240 format with Y component filled with the textures and padding zeros in U and V components. To evaluate the similarity metrics performance, we encoded these videos by replacing the distortion measure

of HM by a scaled version of D_{SSIM} and D_{STSIM} in Eqn. (3) and Eqn. (4).

B. Experiments

We experiment using SSIM and STSIM inside the HM encoder as a distortion measure. Since our videos consist of one frame, these metrics were only tested for Intra-Picture prediction scheme. The distortion measure that we replaced with them has an effect at two computation levels:

1) Hadamard Transformed Sum of Absolute Difference (SATD): This metric is used to select the three most probable intra-prediction modes. These modes, as discussed in Sec. II, will be checked by the encoder in order to find the best rate-distortion tradeoff considering also the partitioning of the prediction blocks.
2) Sum of Square Difference (SSD): The encoder uses this metric at several levels. we replaced its use only for the distortion measure of best mode selection in a particular partition of the prediction blocks

SSIM an STSIM were used in our experiment to compare patches of textures that correspond to the size of the prediction block. For STSIM, we used the simplified design of the steerable pyramid filter developed in [13]. The number of orientations we chose for the frequency decomposition is 2 while the number of scales was varying according to the patch size. We chose the number of scales equals to 2 for the HEVC prediction block sizes of 64x64, 32x32 and 16x16 and 1 for the size of 8x8 and 4x4. the experiments carried out are:

- Experiment 1: We replaced the SATD by the distortion measures of D_{SSIM} and D_{STSIM} from Eqn. (3) and Eqn. (4). We scaled both of them to match the range of SATD as follows:

$$D^1_{SSIM} = D_{SSIM} * 255 * BlockSize$$

$$D^1_{STSIM} = D_{STSIM} * 255 * BlockSize$$

  where 255 is the maximum pixel value for 8 bits integer representation.
- Experiment 2: Similar to the first Experiment, we replaced the SSD by a scaled version of D_{SSIM} and D_{STSIM} which are defined as:

$$D^2_{SSIM} = D_{SSIM} * 255^2 * BlockSize$$

$$D^2_{STSIM} = D_{STSIM} * 255^2 * BlockSize$$

- Experiment 3: In this experiment, we replaced both SATD and SSD by the corresponding SSIM and STSIM measure used in the two previous experiments.

For all of the above experiments, The HM encoder were used to encode the texture videos with different Quantization Parameters (QP). The QP we chose were (22,26,32,36,43,51) to cover a wide range of compression (from fine to coarse compression). We studied the effect of using these metrics on the decoded picture quality. We studied also the effect of using these metrics on the prediction mechanism of HEVC. The results of experiments are given in the next section.

IV. Experimental Results

To show the result of the three experiments on the decoded pictures, we first provide one example of a decoded texture for each of experiments explained before. The effect of using different distortion metrics is not very distinguishable for lightly compressed images, that is why, we show here effect in a very high compression scenario. Fig. 1 shows the effect of replacing SATD with SSIM and STSIM for QP value of 51. It can be seen that replacing SATD with D^1_{SSIM} or D^1_{STSIM} reduces the number of DC blocks and enhances slightly the quality of the decoded image. This has very little improvement because in this approach, the new metrics affect only choosing the 3 most probable intra-prediction modes but doesn't decide on neither the mode selection nor the block partition.

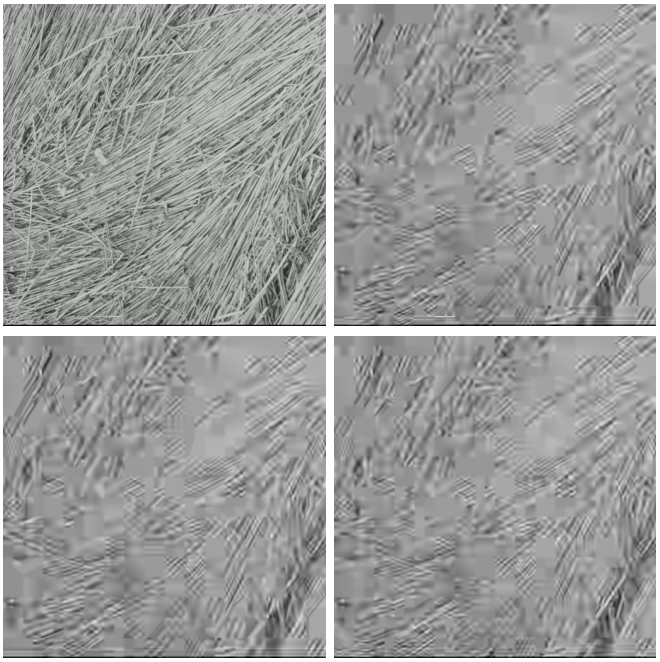

Fig. 1. Effect of replacing of SATD with D^1_{SSIM} and D^1_{STSIM} for QP value 51. Top left: original texture, top right: encoded using default distortion function, down left: SSIM instead SATD, down right: STSIM instead of SATD

Replacing SSD with D^2_{SSIM} or D^2_{SSIM} increases dramatically the quality of the decoded image. this can be seen in Fig. 2. If we look carefully on the decoded textures. It can be seen that STSIM introduces some artificial lines which were not available in the original image. The reason is that STSIM is rotationally invariant metric. With this property, the prediction signal generated using directional prediction may have little distortion computed by STSIM although it is in a wrong direction as compared to the original image. Texture coded with SSIM can maintain better the directionality of the texture (due to the pixel by pixel comparison inherited in it) in some texture blocks, but the overall quality of the texture coded with STSIM looks more natural.

The effect of replacing both SATD and SSD is not very different from replacing SSD only since replacing SATD has a minor effect on the decoded picture as seen before. One

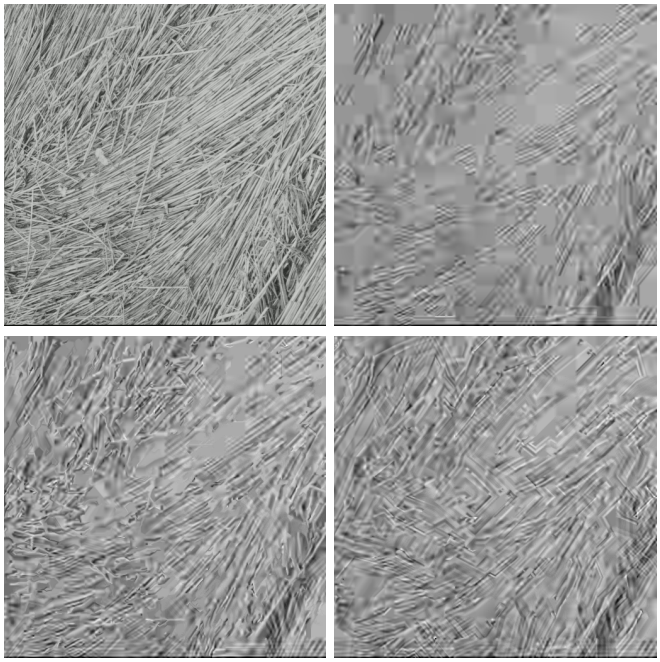

Fig. 2. Effect of replacing SSD with D^2_{SSIM} and D^2_{STSIM}, QP value 51. Top left: Original Image, top right: encoded using default distortion measure, down left: SSIM instead of SSD, down right: STSIM instead of SSD

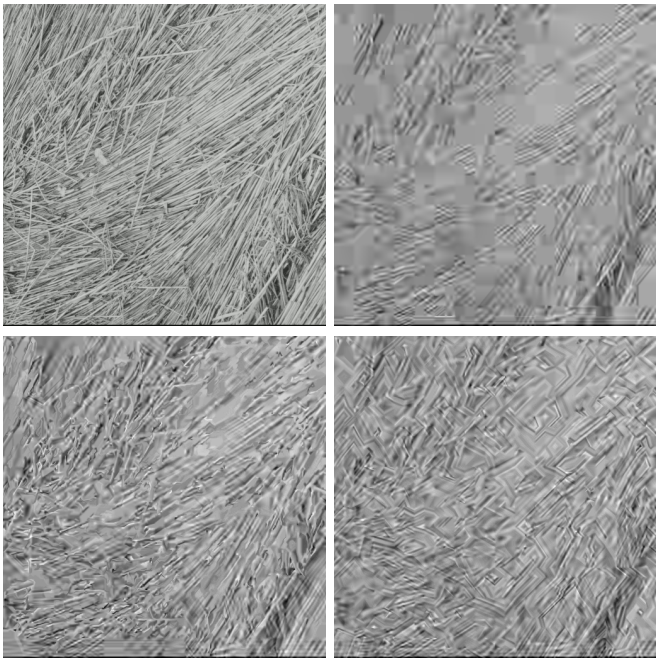

Fig. 3. Effect of replacing both SATD and SSD with SSIM and STSIM, QP value 51. Top left: Original Image, top right: encoded using default distortion measure, down left: SSIM instead of SATD and SSD, down right: STSIM instead of SATD and SSD

thing can be noticed is that the effect of producing artificial lines in STSIM is more obvious here as the all mode selections doesn't relay on pixel by pixel comparison.

Another example is given in Fig. 4 of replacing SATD and SSD with D^2_{SSIM} and D^2_{STSIM} for a non-directional texture. It can be seen that both SSIM and STSIM reduce the amount of visually noisy blocks. It can also be noticed that reconstructed texture looks more natural when using STSIM as compared to SSIM. That is, for a non-directional texture, STSIM performs very well even though some blocks may be predicted in a wrong direction.

To show the effect of these metrics on all of the used texture dataset, Fig. 5 and 6 show the result of coding all textures in third experiment. Looking at these figures, we can draw the same conclusion as before: the quality of the decoded textures increases in using either SSIM or STSIM, textures coded with STSIM look more natural specially for non-directional textures.

To study the effect of each metric on the prediction mechanism of HEVC, we ran the same simulation as before and measured the number of times that the encoder chooses a size of the prediction block in different QP. For our 13 textures, we computed the average value of these numbers and obtained the curves shown in Fig. 7. These curves shows how the encoder works using different distortion metrics.

The first thing that one can notice is the high correlation between the curves of default distortion measure, "SSIM 1"(replacing SATD with D^1_{SSIM}) and "STSIM 1" (replacing SATD with D^1_{STSIM}). The reason behind it is that the behavior of the encoder doesn't change very much in these experiments, SATD only affects selecting the best candidates for rate-distortion optimization. This doesn't have a big influence on the decoded pictures as shown before. It can be even proven if you look at the curves of "SSIM 2" and "SSIM Both" and also the curves of "STSIM 2" and "STSIM Both" where "2" here represents replacing SSD by the corresponding distortion measure and "Both" means replacing both SATD and SSD by them. The very high similarity between them indicates that replacing SATD as well as SSD is not very different from replacing SSD alone.

Another observation from these curves is that the encoder tends to use larger prediction blocks for higher compression. By using SSIM or STSIM, the encoder is forced to use more number of smaller blocks to provide better prediction of the signal.

V. Conclusion

In this paper, we have studied the possibility of perceptual optimization of HEVC for texture components. We used one recent perceptual texture similarity metric, namely, STSIM, as a distortion function for mode selection and block partitioning in intra-prediction scheme of HEVC. This metric was implemented and compared to a the default metrics used in HEVC and also to SSIM which is as well known similarity metric. Both metrics have shown a significant improvement of the quality of decoded textures specially for high compression range.

The implicit pixel by pixel comparison incorporated in SSIM prevents it from selecting wrong direction in the set of

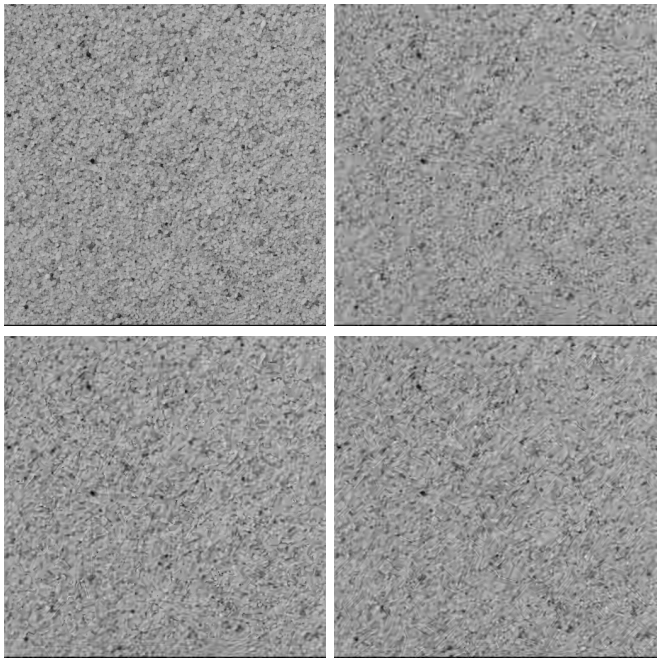

Fig. 4. Effect of replacing both SATD and SSD with SSIM and STSIM, QP value 43. Top left: Original Image, top right: encoded using default distortion measure, down left: SSIM instead of SATD and SSD, down right: STSIM instead of SATD and SSD

directional prediction modes defined in HEVC as happens with STSIM, but generally, the quality of the decoded textures tends to be better although they might differ more in pixel by pixel comparison. In other word, the *non-reference* quality of the decoded textures is improved using STSIM. This observation is more accurate for stochastic textures rather than directional textures.

The rate-quality optimization of the used metrics was out of the scope of this paper. This can be done by appropriately selecting the Lagrangian multiplier of the rate-distortion function. This is left for a possible future research.

VI. ACKNOWLEDGEMENT

This work was supported by the Marie Curie Initial Training Network under the PROVISION (PeRceptually Optimizaed VIdeo CompresSION) project.

REFERENCES

[1] G. J. Sullivan, J. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (hevc) standard," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 22, no. 12, pp. 1649–1668, 2012.

[2] H. R. Wu, A. R. Reibman, W. Lin, F. Pereira, and S. S. Hemami, "Perceptual visual signal compression and transmission," 2013.

[3] J. Balle, A. Stojanovic, and J.-R. Ohm, "Models for static and dynamic texture synthesis in image and video compression," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 5, no. 7, pp. 1353–1365, 2011.

[4] F. Zhang and D. R. Bull, "A parametric framework for video compression using region-based texture models," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 5, no. 7, pp. 1378–1392, 2011.

[5] J. Ehmann, T. Pappas, and D. Neuhoff, "Structural texture similarity metrics for image analysis and retrieval," 2013.

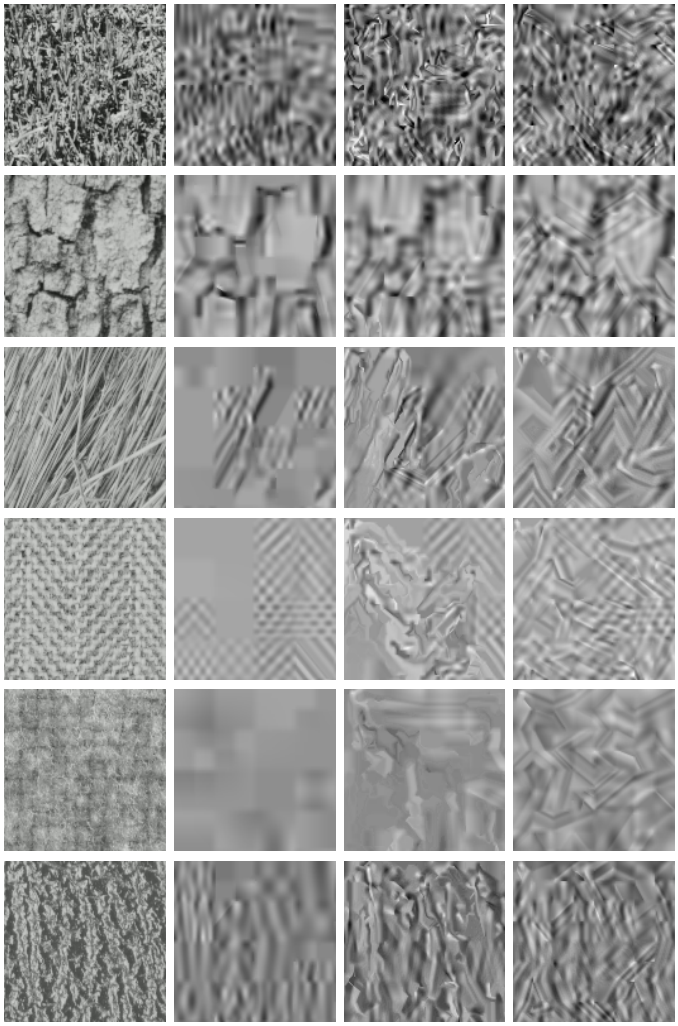

Fig. 5. Effect of replacing both SATD and SSD with SSIM and STSIM, QP value 51. First column: original textures, second column: encoded using default distortion measure, third column: SSIM instead of SATD and SSD, forth column: STSIM instead of SATD and SSD

[6] G. Jin, R. Cohen, A. Vetro, and H. Sun, "Joint perceptually-based intra prediction and quantization for hevc," 2012.

[7] Z. Wang and E. P. Simoncelli, "Translation insensitive image similarity in complex wavelet domain," in *In Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP05). IEEE International Conference on*. Citeseer, 2005.

[8] A. C. Brooks, X. Zhao, and T. N. Pappas, "Structural similarity quality metrics in a coding context: exploring the space of realistic distortions," *Image Processing, IEEE Transactions on*, vol. 17, no. 8, pp. 1261–1273, 2008.

[9] X. Zhao, M. G. Reyes, T. N. Pappas, and D. L. Neuhoff, "Structural texture similarity metrics for retrieval applications," in *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*. IEEE, 2008, pp. 1196–1199.

[10] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger, "Shiftable multiscale transforms," *Information Theory, IEEE Transactions on*, vol. 38, no. 2, pp. 587–607, 1992.

[11] Joint Collaborative Team on Video Coding (JCT-VC) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG, "HEVC test model 9.0," Tech. Rep., 2012.

[12] "Usc viterbi texture dataset." [Online]. Available: http://sipi.usc.edu/database/database.php?volume=textures&image=17

[13] K. Castleman, M. Schulze, and Q. Wu, "Simplified design of steerable

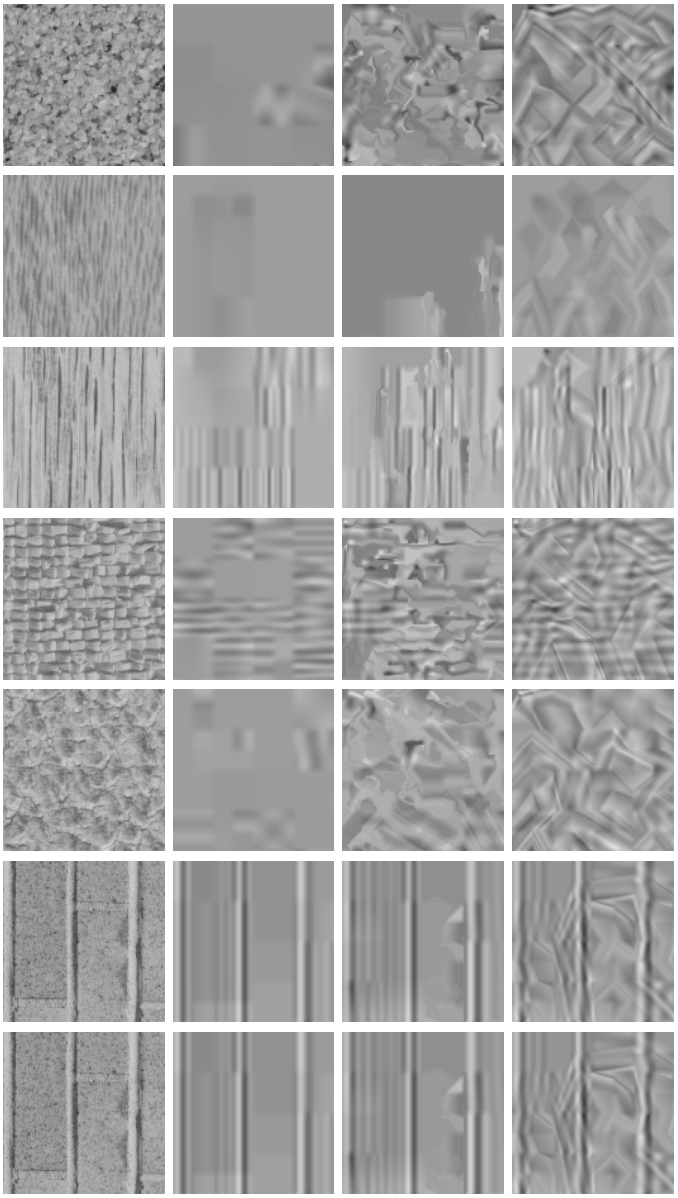

Fig. 6. Effect of replacing both SATD and SSD with SSIM and STSIM, QP value 51. First column: original textures, second column: encoded using default distortion measure, third column: SSIM instead of SATD and SSD, forth column: STSIM instead of SATD and SSD

pyramid filters," in *IEEE International Symposium on Circuits and Systems*. INSTITUTE OF ELECTRICAL ENGINEERS INC (IEEE), 1998, pp. 329–332.

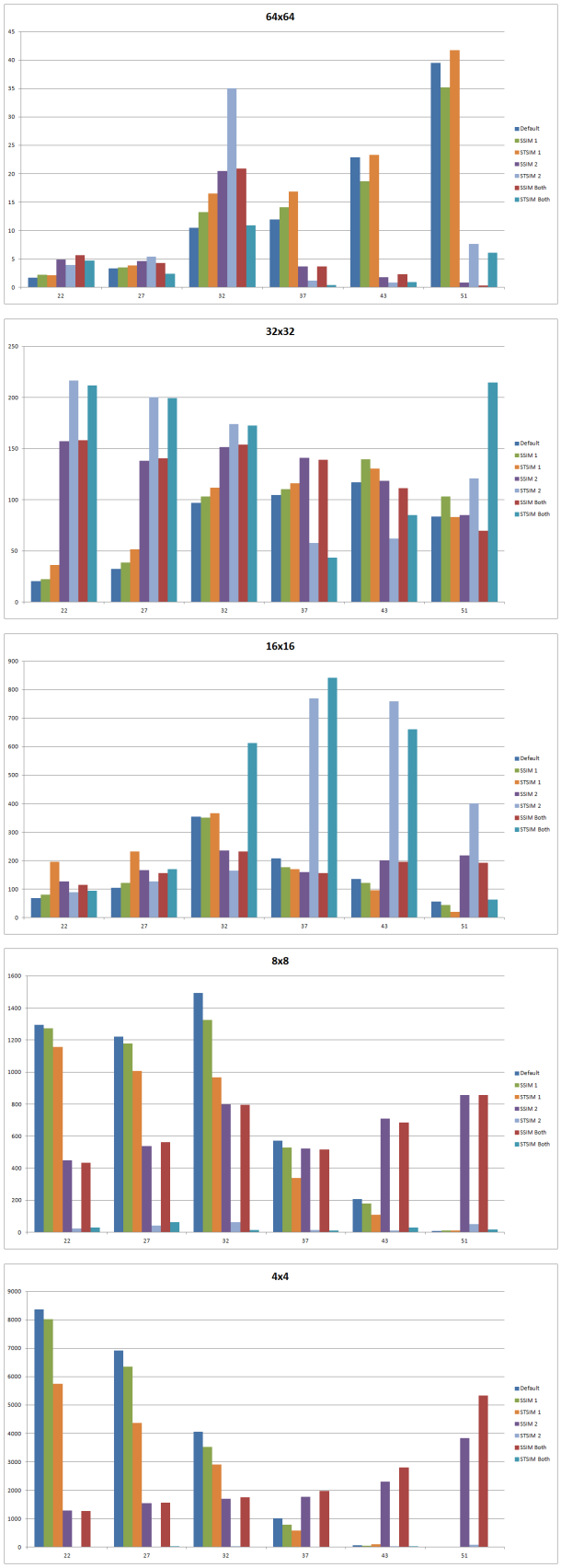

Fig. 7. Average number of times of selecting a prediction blocks size as a function of the Quantization parameter. "SSIM 1" and "STSIM 1": replacing SATD with SSIM and STSIM resp., "SSIM 2" and "STSIM 2": replacing SSD with SSIM and STSIM resp., "SSIM Both" and "STSIM Both": replacing both SATD and SSD with SSIM and STSIM resp., horizontal axes: QP, vertical axes: number of time